

EXHIBIT 147

UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, et al.,

Individual and Representative Plaintiffs,

v.

META PLATFORMS, INC.,

Defendant.

Case No. 3:23-cv-03417-VC

OPENING EXPERT REPORT OF
CRISTINA VIDEIRA LOPES, PhD

JANUARY 10, 2025

works in the Common Crawl would require computational resources and time that go beyond the constraints of my assignment.

145. In order to search for the Plaintiffs' works in Meta's training data, I was given a collection of hard (in most cases) and electronic (in a few cases) copies of Plaintiffs' books. I then extracted text excerpts from the beginning, middle, and end of each book, and proceeded to search whether those text excerpts were found in B3G, LIBGEN11, and LIBGEN12. My detailed findings are shown in Appendix F.

146. B3G, LIBGEN11, and LIBGEN12 are very large collections of data. As such, I expedited my search with scripts that created and used a "table of contents" for each dataset, making searching considerably faster than a naïve search over the entire dataset. What follows describes the datasets with more detail, the methods I used for creating and using tables of contents, and, finally, a discussion of my findings.

E. BOOKS DATASETS

147. B3G contains 139,588 entries from the books3 dataset; LIBGEN11 contains 758,244 entries from LibGen, and LIBGEN12 contains 711,451 entries from LibGen. Each "entry" is, roughly, a copy of a book, although I found cases where one "entry" included copies of two or more books attached to each other. In any case, one "entry" corresponds to one single line in the dataset files.

148. As an illustration, the following is the first line (aka entry) of the first chunk file of B3G:

```
{"footer_strip_count":1,"header_strip_count":1,"lang_pred_books3":[["__label__en","__label__de","__label__ru"],["0.9735388159751892","0.0018873148364946246","0.001547140534967184"]], "path":"hdfs://ha-nn-uri/data/books3/raw/2022-09-18/S/slaveshiptheredikermarcus.epub.txt", "source":"books3", "text":"Lying in the bottom of the canoe ..."}

```

149. The example above shows a copy of the book "The Slave Ship" by author Marcus Rediker. The title and author information are in the "path" field, with all words concatenated ("slaveshiptheredikermarcus"); the origin of this data is in the "source" field, in this case Books3; the text of the book itself is in the "text" field – the example above shows only the first few words of the book, but the data file contains the complete copy of the book. The field

“**lang_pred_books3**” is the result of some language detection classifier, in this case showing that the language is English.

150. The fields “**footer_strip_count**” and “**header_strip_count**” are likely data included by Meta indicating whether lines from the copies of the book were removed from the beginning and the end of those copies. In this case, it looks like one line was removed from the header and another from the footer. Very likely the removed lines were related to copyright notices present in the text of the book. Unfortunately, the script used to create the B3G collection was not made available to me, so I cannot assert with certainty that the removed lines were related to copyrights. But given other fields in the LIBGEN datasets (see next), as well as internal documents produced by Meta,²⁷ it is more likely than not that this is the case.

151. The two LibGen datasets have different information from B3G. As an illustration, the following shows the first line of the first chunk file of LIBGEN11:

```
{"source": "f3fe976e42e5044210f29af5f55c7c0c.md", "format": "epub", "text":  
"For my Lovelies...", "filter": {"lines_copyright_removed": 0,  
"newlines_removed": 3922, "lines_pii_removed": 1, "lines_repetition_removed":  
0}}
```

152. Similarly, the following shows the first line of the first chunk file of LIBGEN12:

```
{"filter":{"lines_copyright_removed":0,"lines_pii_removed":0,"lines_repetitio  
n_removed":0,"newlines_removed":3229},"format":"epub","source":"58553ec76879d  
cab2253c3af6545cbc4.md","text":"# Chapter One...",  
"scores":{"nsfw_ident_v1":0.004153358909951628}}
```

153. Although the fields are in a different order, the fields of the entries of LIBGEN11 and LIBGEN12 are almost identical, with LIBGEN12 having one additional field “**nsfw_ident_v1**”, in this case having a very low 0.0041 value. NSFW typically means “not suited for work,” an expression used to denote pornography and other undesirable types of content. As far as I can tell, this is the only difference between LIBGEN11 and LIBGEN12, suggesting that Meta performed some sort of content filtering to the December version of LibGen.

²⁷ See, e.g., Meta_Kadrey_0054898 at -899 (Meta employee Melanie Kambadur stating “sometimes we do move the copyright text as part of common cleaning because it is very repeated across datasets” in connection with a discussion about Books3).

154. Importantly, both LIBGEN11 and LIBGEN12 include fields related to the removal of copyright (“**lines_copyright_removed**”) and personal identifiable information (“**lines_pii_removed**”). Similarly to B3G, the text of the books is in the field “**text**,” with the two examples above showing only the first few words.

155. Unlike with B3G, there is no author or title information in the LibGen datasets present in the hard drives made available to me. Each entry/line contains just the text of books without any attribution or association with the author. This made my searching process considerably more difficult, especially considering that Meta has in its possession the complete information of the books in LibGen in the form of a relational database (the catalogue). That database was not made available to me.

F. METHODS

156. The datasets produced by Meta in this case are voluminous. In order to more manageably search and analyze each of the datasets for my assignment, I first built tables of contents (TOCs) for B3G, LIBGEN11, and LIBGEN12, then I searched. The scripts I used to create these TOCs and perform search are shown in Appendix E.

157. TOCs speed up the searches considerably because they allow me to first search for relatively small data (authors, titles, beginning/ending of books) and then perform full text search only on the entries that match the small data search.

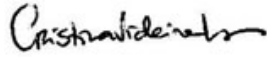
158. For B3G, the TOC entries look like the following example:

17, hdfs://ha-nn-uri/data/books3/raw/2022-09-18/L/letterofmarya
laurierking.epub.txt, b3g.chunk.00.jsonl, 473009, 12696311

159. In the above example, “17” denotes the line number in the chunk file; the part starting with hdfs is the “path” containing the author and title information; b3g.chunk.00.jsonl is the chunk file where the entry is; 473009 is the size of the text in characters; and 12696311 is the offset of this entry in the chunk file (in bytes).

160. For the LibGen datasets, the TOC entries are more verbose because of the lack of author and title information in the data files. LibGen TOCs look like the following example:

Respectfully Submitted,

A handwritten signature in black ink, appearing to read "Cristina Videira Lopes", written over a horizontal line.

Cristina (“Crista”) Videira Lopes, PhD

January 10, 2025